**Pattern Recognition Project (PHASE 1)**

**Game Popularity**

**Team ( CS\_26 )**

|  |  |  |
| --- | --- | --- |
| المستوي | رقم الجلوس | الاسم |
| 3 | **20201700383** | **سيف الدين سيد ابراهيم سيد عمران** |
| 3 | **20201700393** | **سيف عز الدين احمد السيد** |
| 3 | **20201700390** | **سيف الدين وائل السيد نصر حجاب** |
| 3 | **20201700382** | **سيف الدين سامر عصمت** |
| 3 | **20191700317** | **شوكت سيد محمد** |

**Pre-Processing.**

We need to apply all preprocessing techniques listed below: -

* Handling Null Values (Fill and Drop).
* Remove Duplicates
* Encode Dates to be numerical value.
* Get useful value from list of numerical values.
* Get useful numerical information from list of non-numerical values.
* Handling Categorical Data.
* Removing additional characters from numerical values.
* Normalization.
* Drop columns with most records equals zero.
* Drop columns that are not useful for the model (Unique Values).

**How do we drop columns or rows?**

We drop rows or columns based on certain conditions

For columns: -

* Contain null values > 30%
* Most of cells equals 0 so it may be confuse the model
* Columns of Unique Values > 90%

For rows: -

* If it contains null values

**Pre-Processing for each feature.**

|  |  |
| --- | --- |
| Feature | Pre-Processing |
| URL | Drop -> Unique Values |
| ID | Drop -> Unique Values |
| Name | Drop -> Unique Values |
| Subtitle | Drop -> Contain null values > 30% |
| Icon URL | Drop -> Unique Values |
| User Rating Count | Normalization |
| Price | Fill null with 0 (free games), Normalization |
| In-app Purchases | Get the average of each list of purchases and fill empty with 0, Normalization |
| Description | Drop |
| Developer | Categorical data converted to dummy variables |
| Age Rating | Remove “+” sign, Normalization |
| Languages | Count number of languages for each game, Normalization |
| Size | Normalization |
| Primary Genres | Categorical data converted to dummy variables |
| Genres | Calculate total number of generes for each game, Normalization |
| Original Release Date | Get the year of release |
| Current Version Release Date | Get the year of release |
| Game Age (new generated) | How many years passed on game release |

**Feature Selection.**

Some features are contains unique values which is not useful for the model and it may confuse the model and increase error such as: -

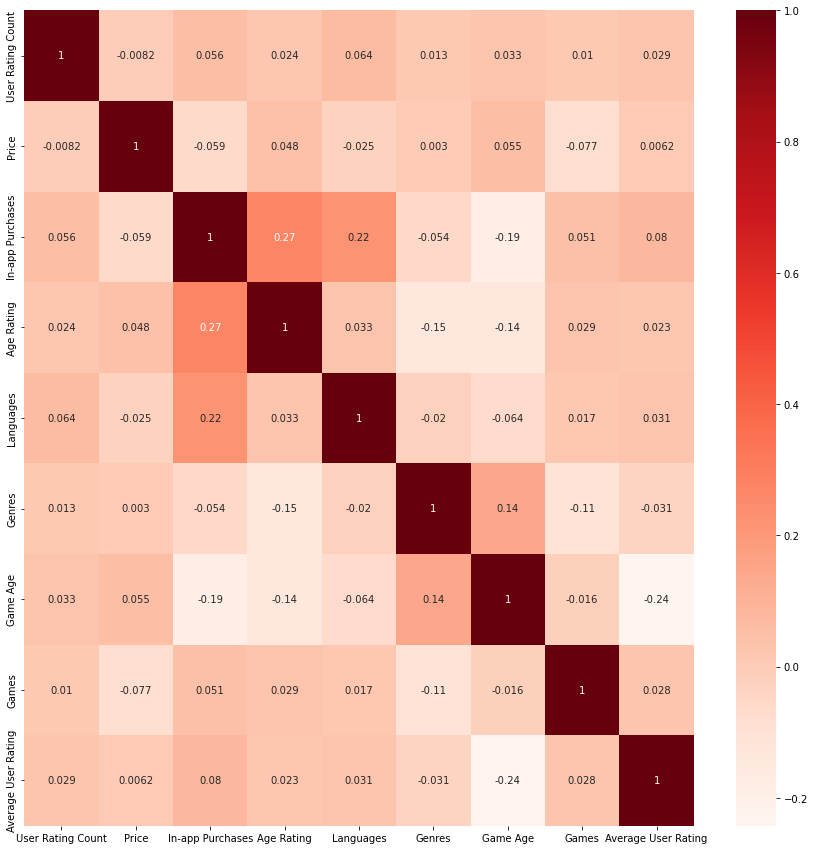
* URL
* Name
* Icon URL
* ID

So, we decided to drop them.

Some other features contains 0 in more than 90% of cells so it may confuse the model as well, Most of these features are the new dummy variables generated from categorical data

So, we decided to drop features with 0 values more than 90%.

**Correlation.**



Based on the correlation matrix we decided to remove features with very low correlation with average user rating.

**Final Used Features**

* User Rating Count
* Age Rating
* In-app Purchases
* Number of Languages
* Size
* Game Age
* Genres
* Current Version Release Date

**Splitting the data.**

We decided to split the data before running any pre-processing techniques.

We decided to use 80:20 % for train : test datasets because the data size is not large enough.

**Models**

We used 3 regression models: -

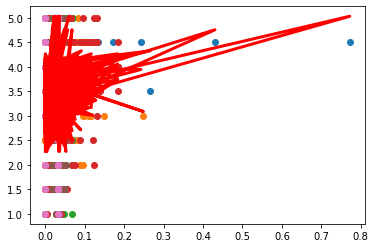
* Multiple Linear Regression
* Ridge Regression
* Lasso Regression

**Model Accuracy**

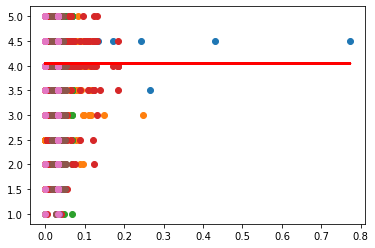
|  |  |
| --- | --- |
| Model Name | Mean Square Error |
| Multiple Linear Regression | 0.84 |
| Ridge Regression | 0.58 |
| Lasso Regression | 0.56 |

**Model Plots**

Linear Regression.

****

Lasso Regression



Ridge Regression

